Boston Housing Dataset Analysis

We took Boston housing dataset from Kaggle website. This dataset has 506 rows and 14 columns. The variable in Boston housing dataset are

- ‘crim’: per capita crime rate by town.
- ‘zn’: proportion of residential land zoned for lots over 25,000 sq.ft.
- ‘indus’: proportion of non-retail business acres per town.
- ‘chas’:Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- ‘nox’: nitrogen oxides concentration (parts per 10 million).
- ‘rm’: average number of rooms per dwelling.
- ‘age’: proportion of owner-occupied units built prior to 1940.
- ‘dis’: weighted mean of distances to five Boston employment centres.
- ‘rad’: index of accessibility to radial highways.
- ‘tax’: full-value property-tax rate per $10,000.
- ‘ptratio’: pupil-teacher ratio by town
- ‘black’: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
- ‘lstat’: lower status of the population (percent).
- ‘medv’: median value of owner-occupied homes in $1000s

```
b <- read.csv("housing.csv")
boston <- b[,-16:-19]
```

- ## Summary of boston housing dataset

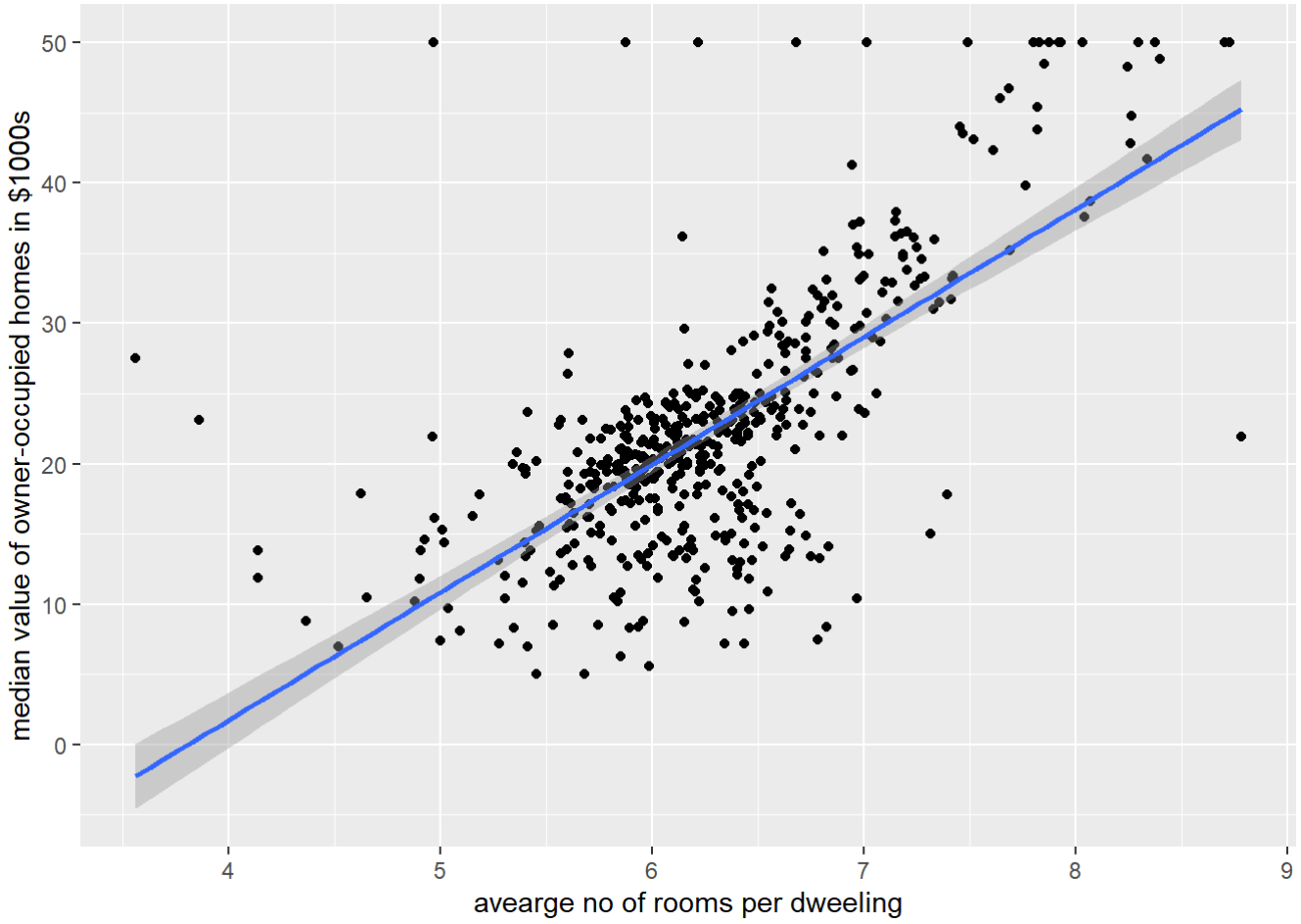
```
summary(boston)
```

```
##        ID              crim                zn              indus       
##  Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46  
##  1st Qu.:127.2   1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19  
##  Median :253.5   Median : 0.25651   Median :  0.00   Median : 9.69  
##  Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14  
##  3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10  
##  Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74  
##       chas              nox               rm              age        
##  Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
##  1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
##  Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
##  Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
##  3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
##  Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
##       dis              rad              tax            ptratio      
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60  
##  1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40  
##  Median : 3.207   Median : 5.000   Median :330.0   Median :19.05  
##  Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46  
##  3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20  
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00  
##      black            Istat            medv      
##  Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
##  1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
##  Median :391.44   Median :11.36   Median :21.20  
##  Mean   :356.67   Mean   :12.65   Mean   :22.53  
##  3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :396.90   Max.   :37.97   Max.   :50.00
```

- Scatter plot between the ages and median value

```
library(ggplot2)
p <- ggplot(boston,aes(x=rm,y=medv))+geom_point()+geom_smooth(method="lm")+labs(x="avearge no of rooms per dweeling",y="median value of owner-occupied homes in $1000s")
p
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
cor(boston$rm,boston$medv)
```

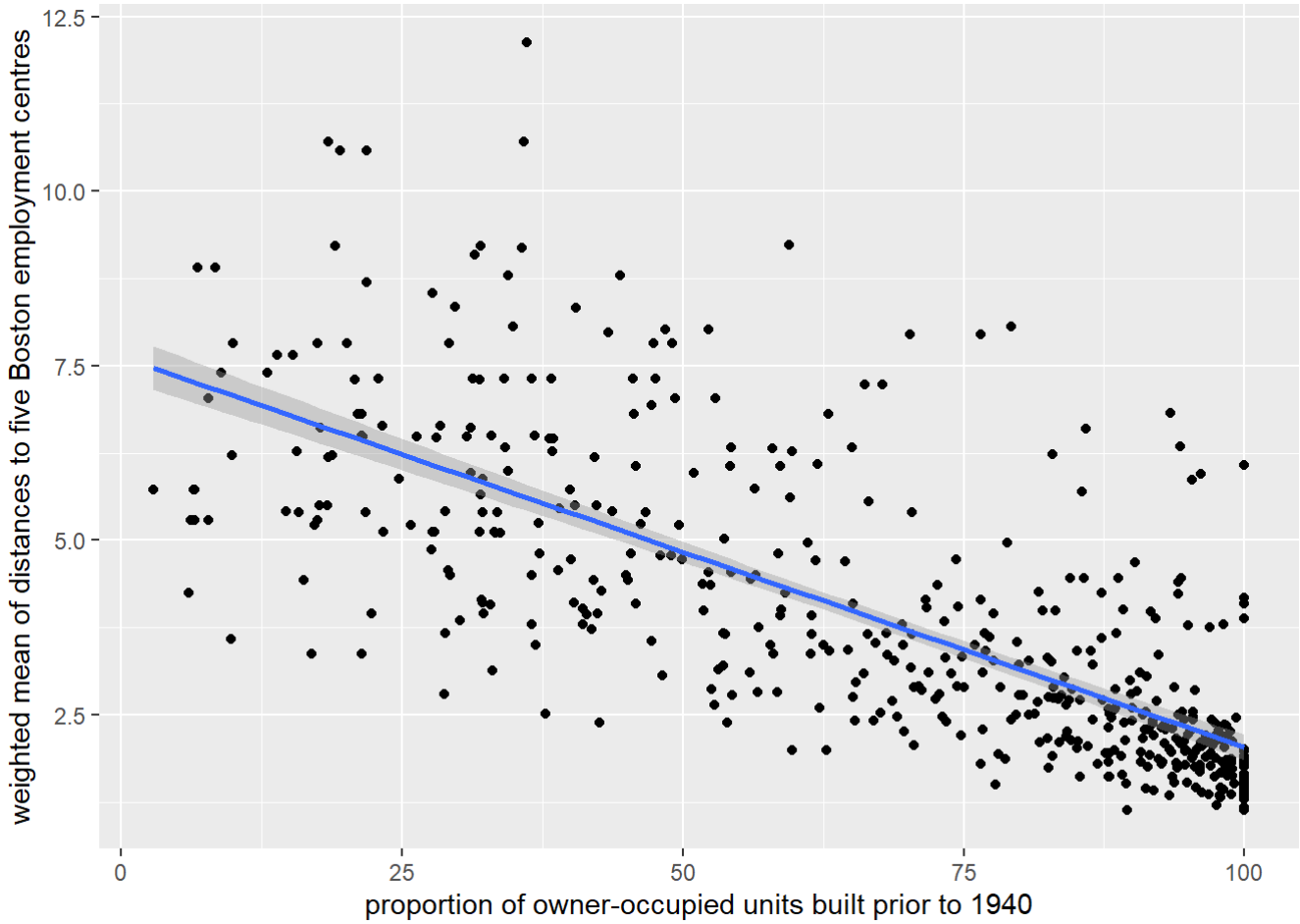
```
## [1] 0.6953599
```

For here, we concluded that the correlation between median value and average no of rooms is positive

- ## Scatter plot between age and distance from five boston employment centre

```
q <- ggplot(boston,aes(x=age,y=dis))+geom_point()+geom_smooth(method="lm")+labs(x="proportion of owner-occupied units built prior to 1940",y="weighted mean of distances to five Boston employment centres")
q
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
cor(boston$age,boston$dis)
```

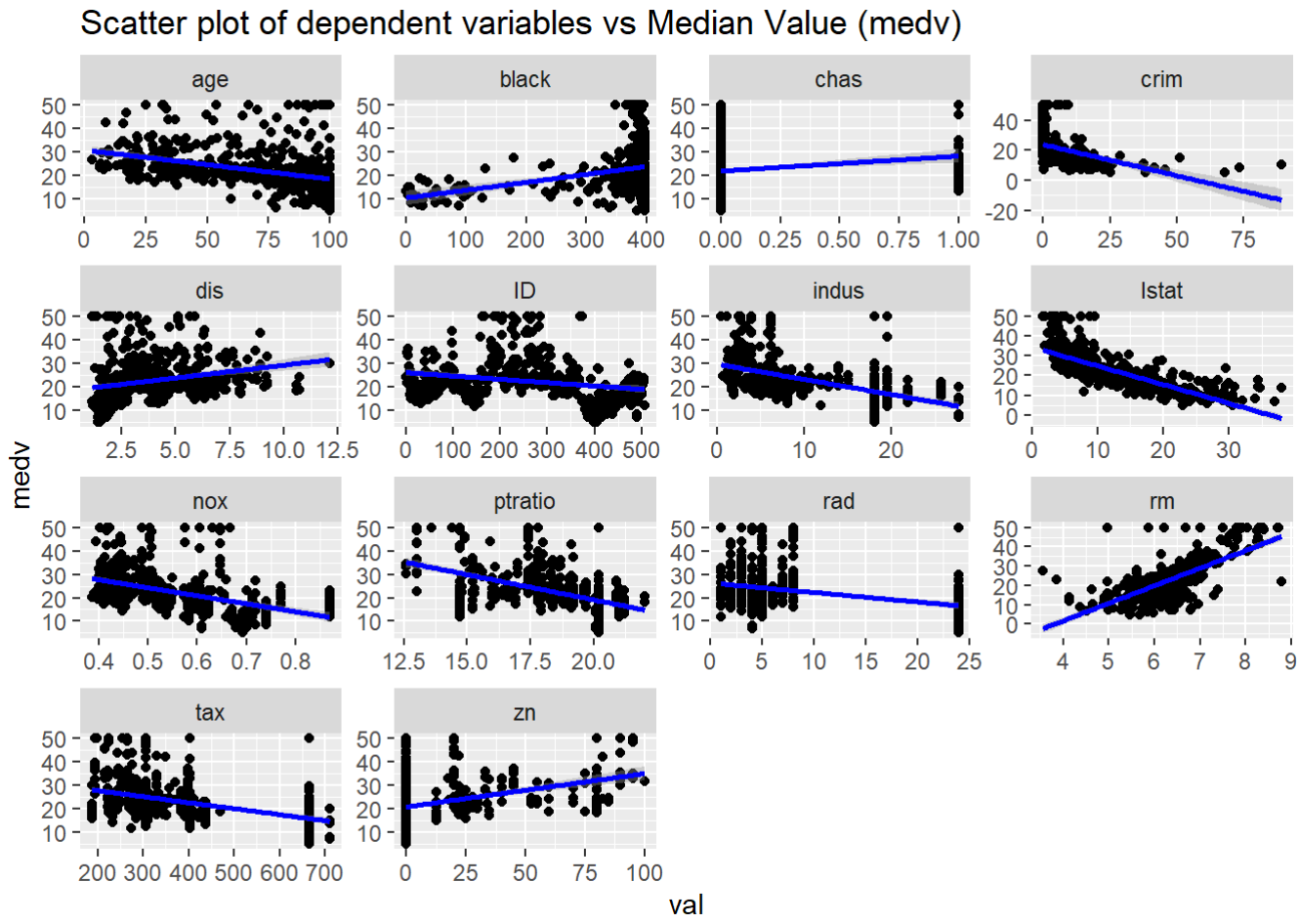
```
## [1] -0.7478805
```

Here the correlation between age and dis is negative,it may be assumed that old-age people take their houses nearer to five boston employment centre,the more aged-people,the lesser the distance to travel to workplace

- # Overall comparsion between all variables of boston housing dataset and median value

```
library(tidyr)
library(ggpubr)
boston %>%
  gather(key, val, -medv) %>%
  ggplot(aes(x = val, y = medv)) +
  geom_point() +
  stat_smooth(method = "lm", se = TRUE, col = "blue") +
  facet_wrap(~key, scales = "free") +
  theme_gray() +
  ggtitle("Scatter plot of dependent variables vs Median Value (medv)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



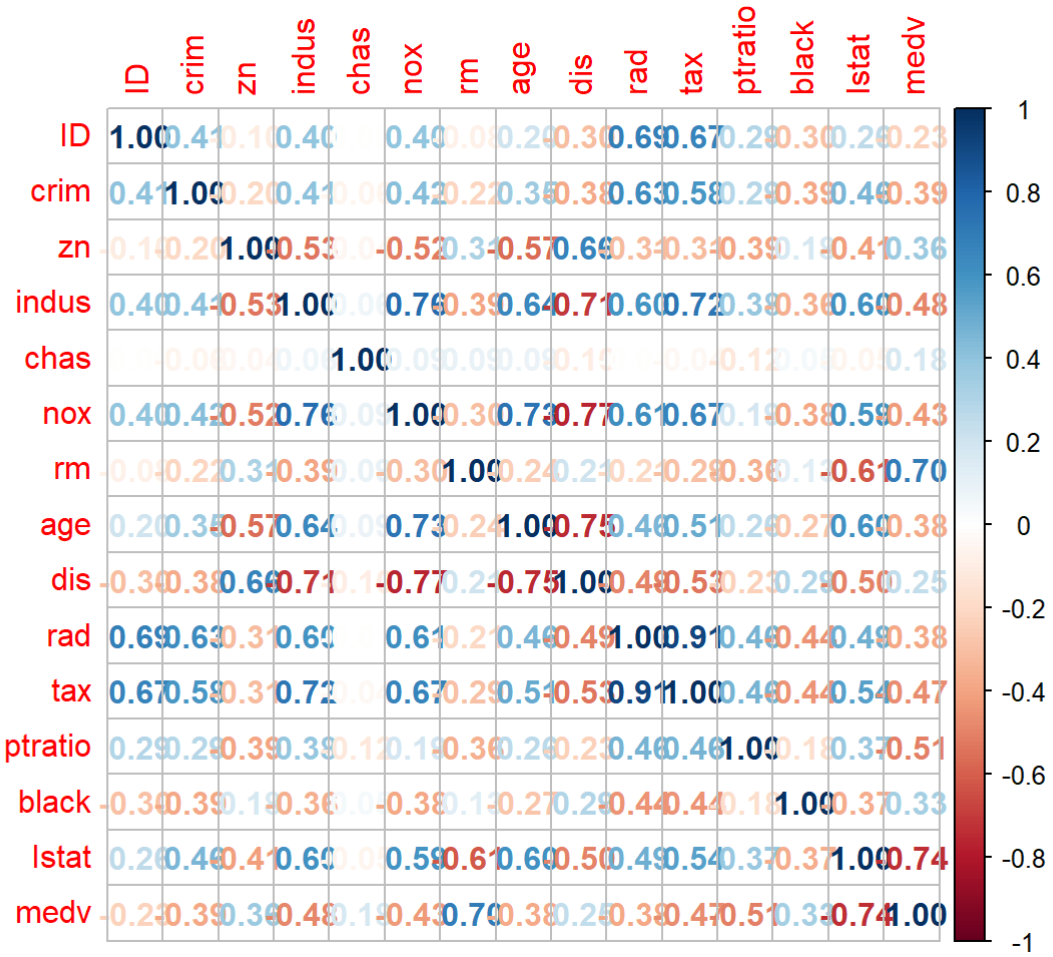
Median value of boston housing dataset depends on other variables as median value increases, as age decreases,black increases,chas increases,crim increases,dis increases,indus decreases, lstat decreases,nox decreases,ptratio decreases,rad decreases,rm increases,tax decreases,zn increases.

- # Correlation matrix

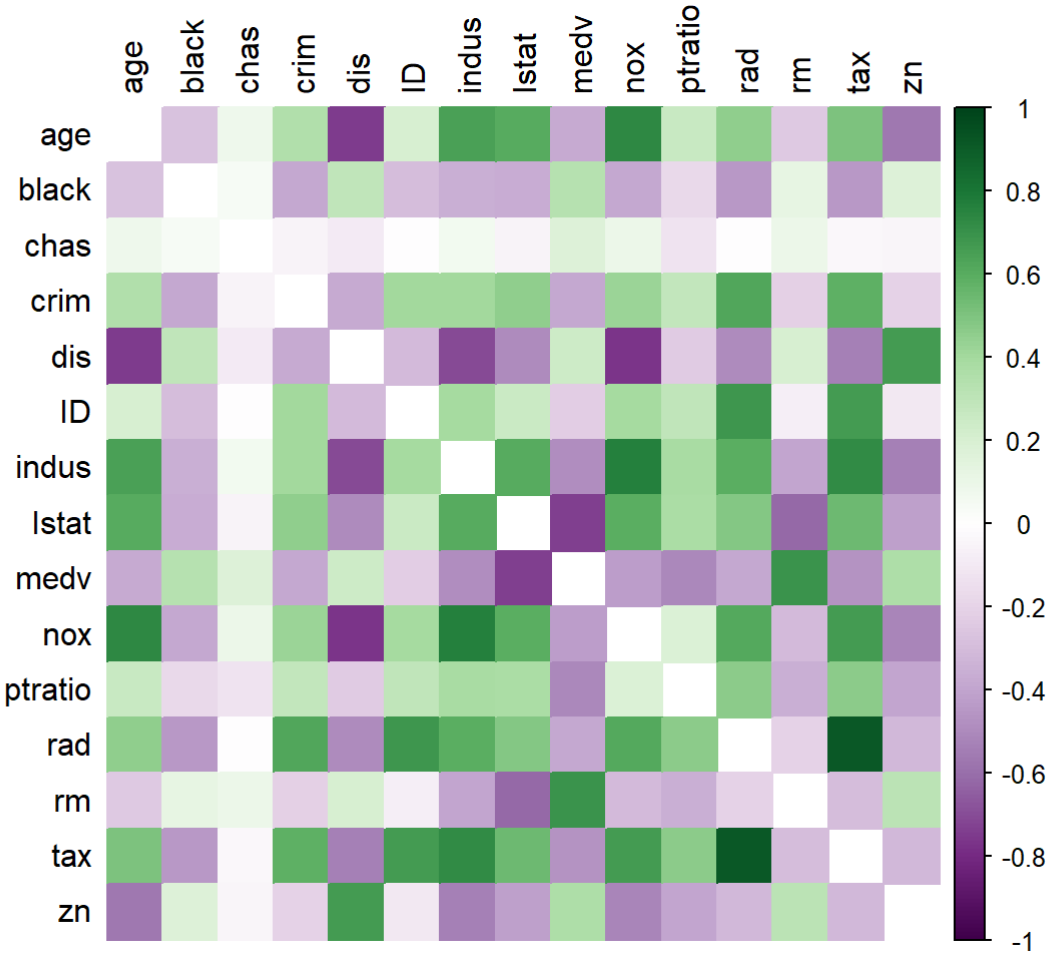
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
j <- cor(boston)
corrplot(j,method='number')
```



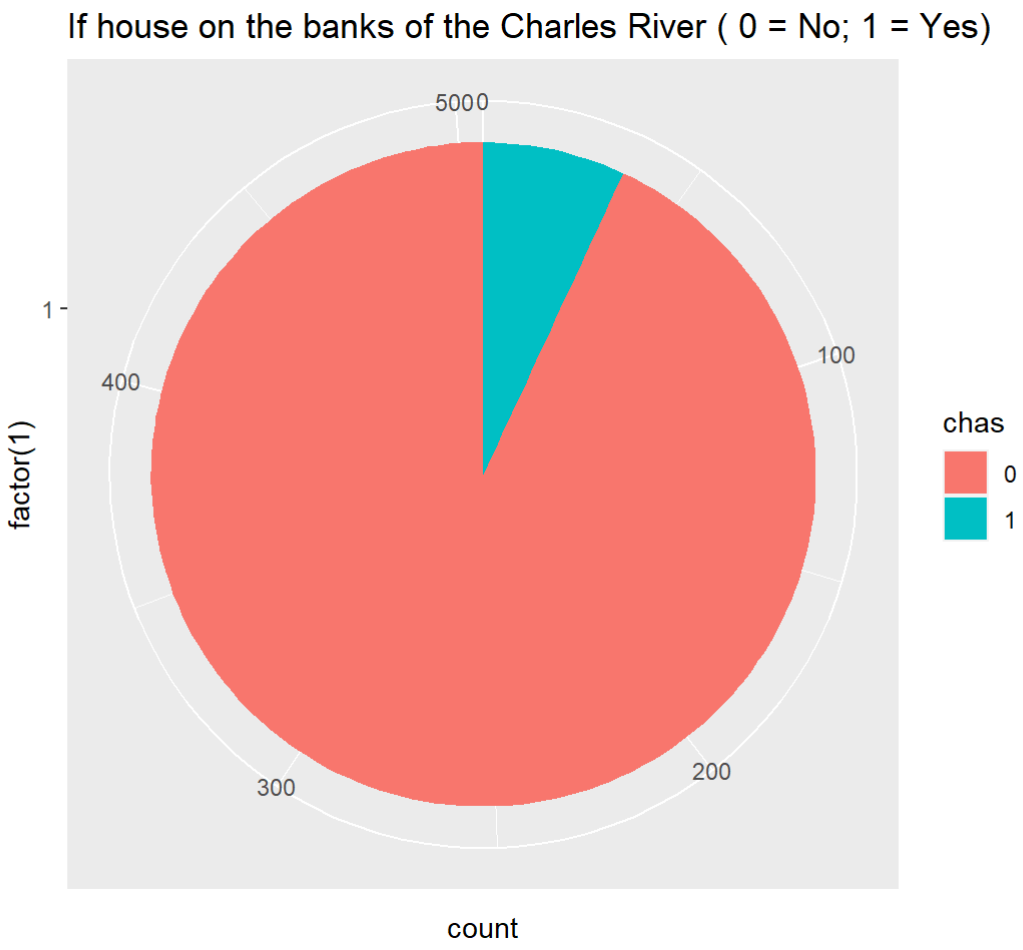

```
corrplot(j,method = 'color', order = 'alphabet',col=COL2('PRGn'),diag=FALSE,tl.col = 'black')
```



- Strong positive coorelation, as the number of rooms increase/decrease, the housing price increases/decreases
- Strong negative coorelation, the more/less the population consists of lower status individuals,housing price decreases/increases.
- Strong negative coorelation, the more/less concentrated NOX is in the air, the lower/higher the price of housing.
- Strong negative coorelation, the more/less concentrated NOX is in the air, the lower/higher the price of housing.
- As the number of students increases for every teacher, the value of housing decreases.
- As them crime rate decreases/increases, the housing price increases/decreases.

Pie chart of Charles River

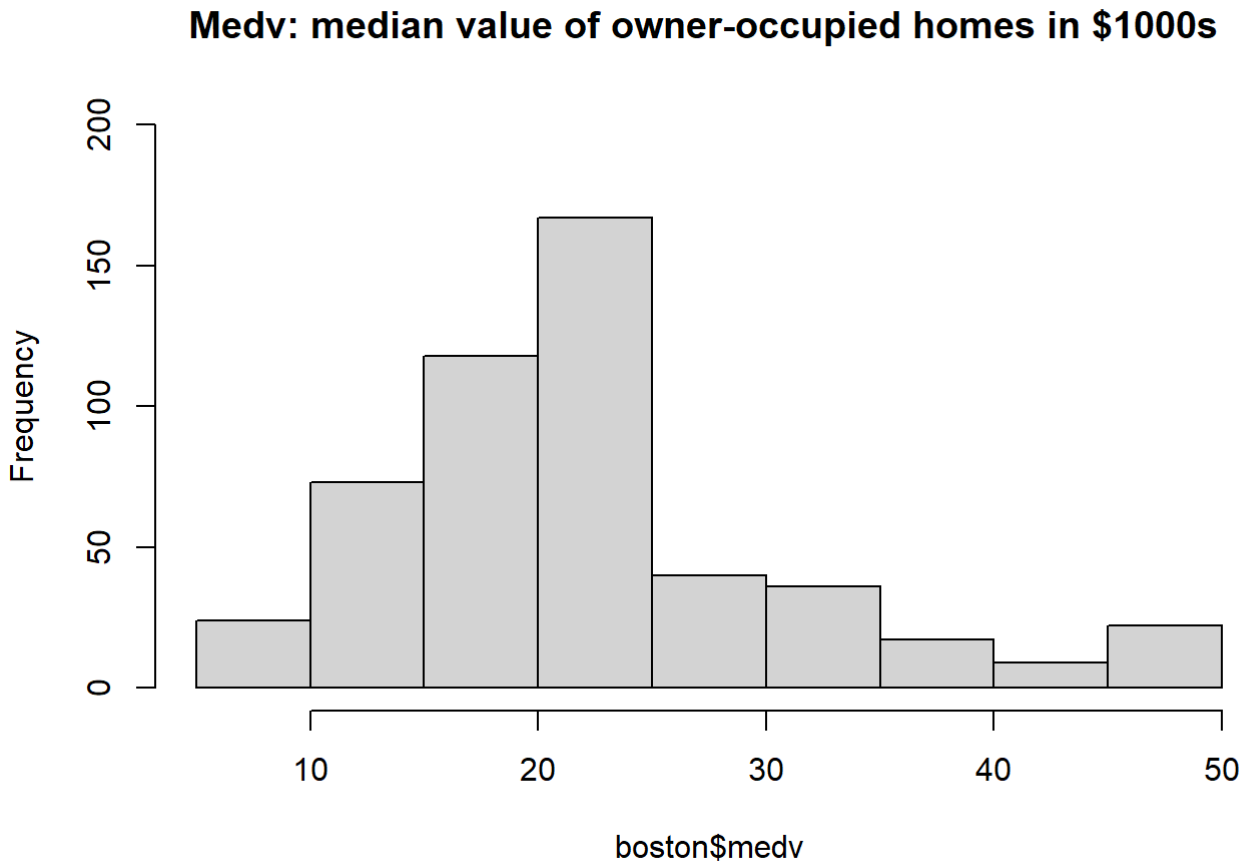
```
riv <- ggplot(boston, aes(x = factor(1), fill = as.factor(chas))) + geom_bar(stat = "count")
 +
  coord_polar("y") + labs(fill = "chas", title = "If house on the banks of the Charles River
 ( 0 = No; 1 = Yes)")
riv
```



Here we can observe that very less people are prefered to live near Charles river.

Histogram of Median value of Boston housing dataset

```
hist(boston$medv,main="Medv: median value of owner-occupied homes in $1000s ",ylim=c(0,200))
```

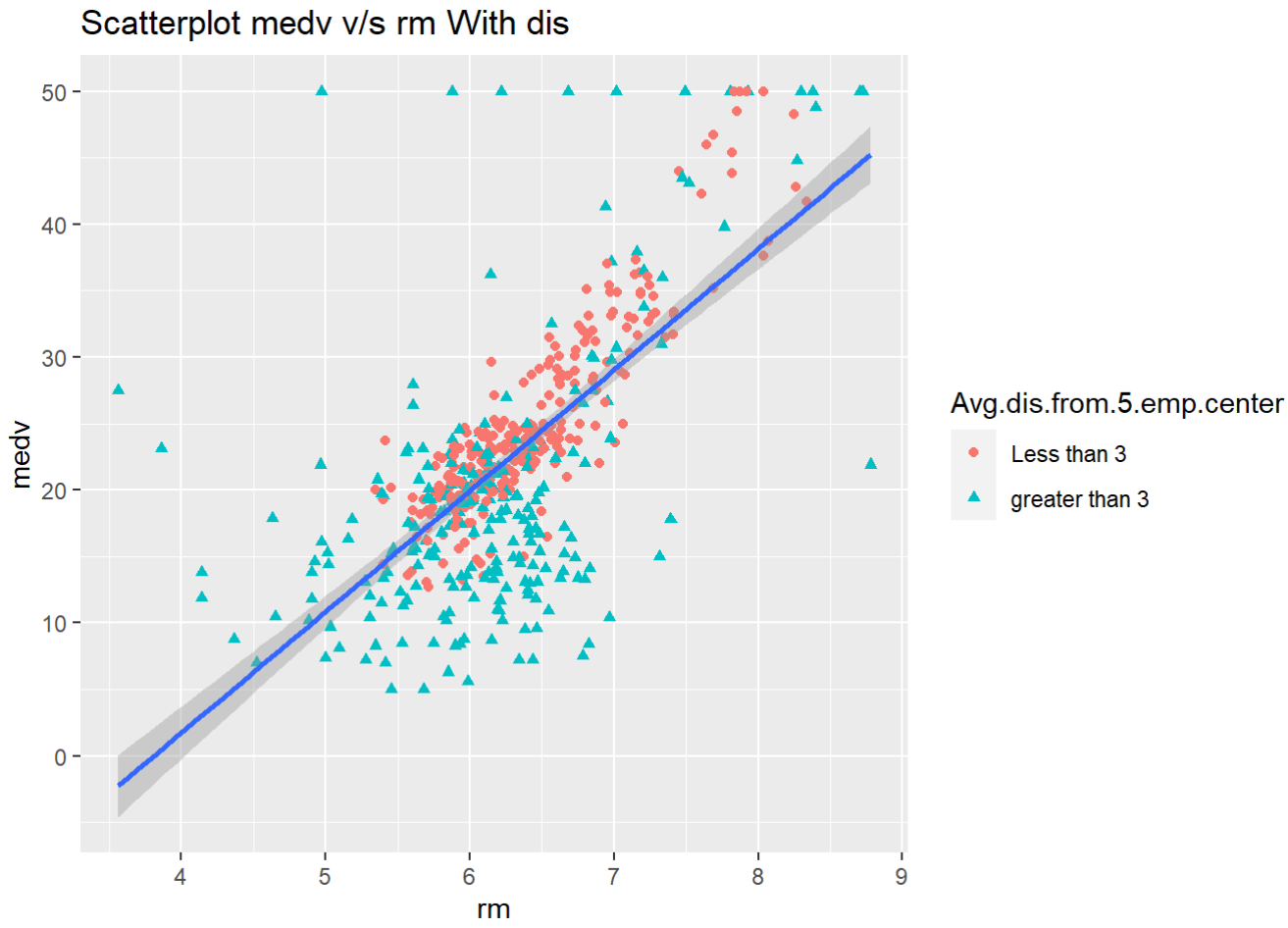


Here we can observe that in histogram of median value around 10 to 20 of value of owner-occupied homes in $1000s have more frequency.The histogram is also left-skewed distribution and mean value is around the beginning or at the end of the data range.

- Some comparative scatter plots
- Scatterplot medv v/s rm with dis

```
Avg.dis.from.5.emp.center <- boston$dis  < 3
Avg.dis.from.5.emp.center <- as.factor(Avg.dis.from.5.emp.center)
levels(Avg.dis.from.5.emp.center) <- c("Less than 3", "greater than 3")
g <- ggplot(boston, aes(rm, medv))+geom_point(aes(shape = Avg.dis.from.5.emp.center, col = Av
g.dis.from.5.emp.center)) +geom_smooth(method = "lm") +labs(x = "rm", y = "medv", title = "Sc
atterplot medv v/s rm With dis")
g
```

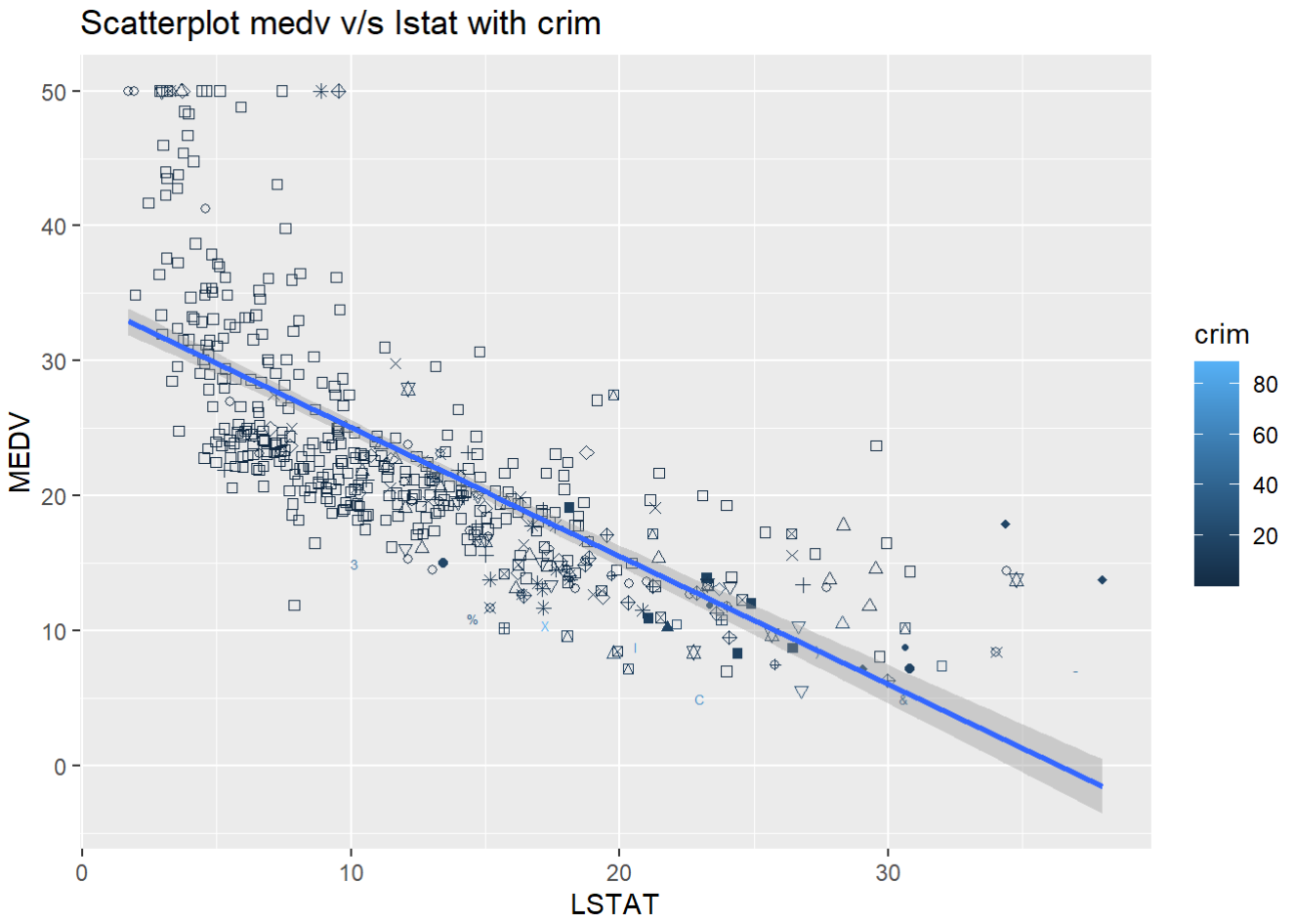
```
## `geom_smooth()` using formula = 'y ~ x'
```



Here we concluded that for a higher rm, one would expect to observe a higher medv.This is because more rooms would imply more space, thereby costing more, taking all other factors constant.Higher priced House are nearer to 5 Boston Emploment center(dis).

- Scatterplot medv Vs Istat With crim

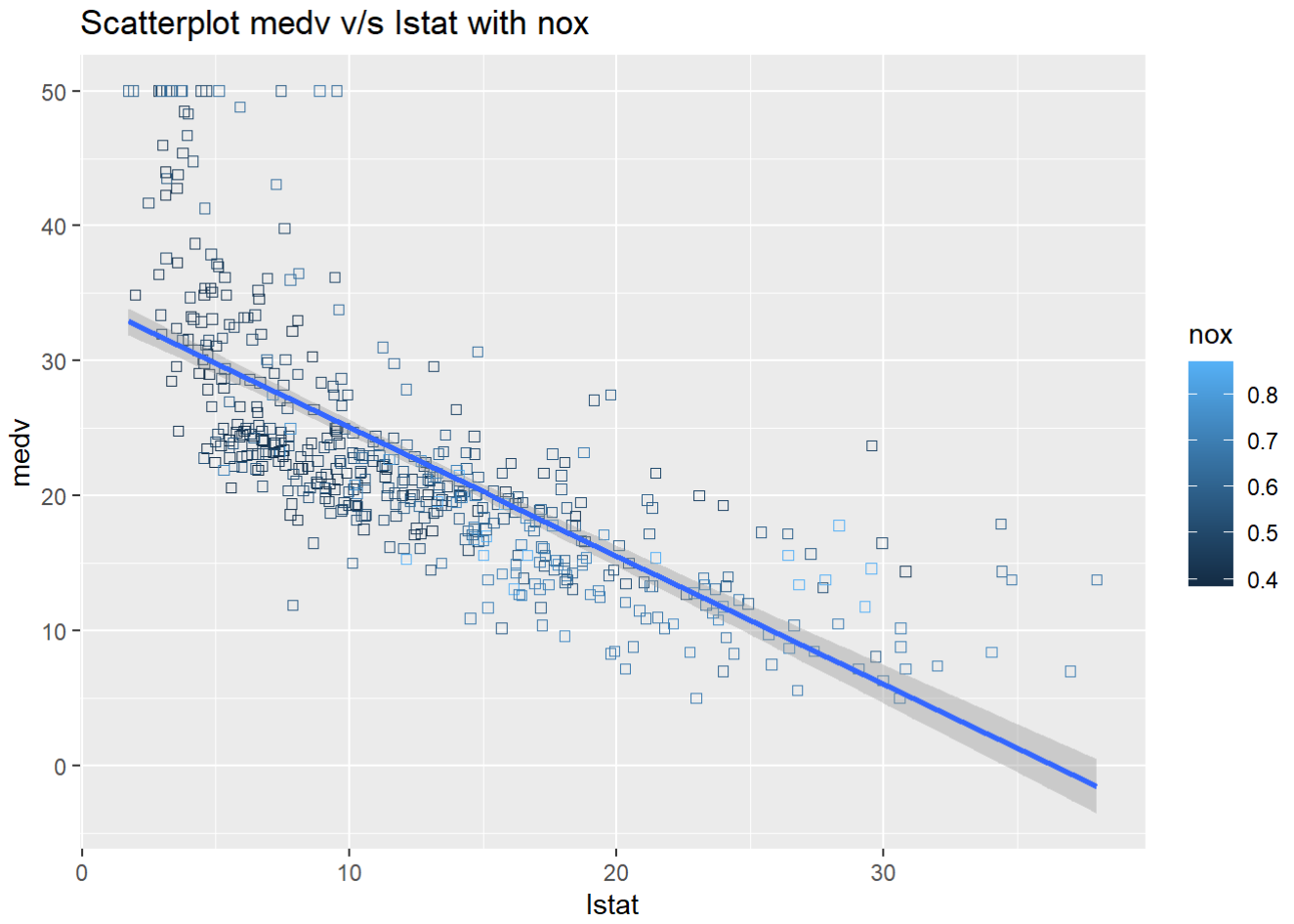
```
crim <- boston$crim  < 20
crim <- as.factor(crim)
levels(crim) <- c("Less than 20", "greater than 20")
g2 <- ggplot(boston, aes(Istat, medv)) + geom_point(aes(shape = crim,col = crim))+geom_smooth
(formula = y ~ x,method = "lm") + labs(x = "LSTAT", y = "MEDV", title = "Scatterplot medv v/s
Istat with crim")+scale_shape_identity()
g2
```



Here we concluded that for a higher LSTAT, one would expect to observe a a lower medv.The social milieux in an area dominated by “lower class” citizens may not be conducive for young children. It may also be relatively unsafe compared to an area dominated by “upper class” citizens. Hence an area with more “lower class” citizens would lower demand, hence lower prices.Crim rate in lower class area is also less.

- Scatterplot medv Vs lstat With nox

```
nox <- boston$nox  < 0.5
nox <- as.factor(nox)
levels(nox) <- c("Less than 0.5", "greater than 0.5")
g3 <- ggplot(boston, aes(Istat,medv))+geom_point(aes(shape = nox,col = nox))+geom_smooth(form
ula= y ~ x,method = "lm") + labs(x = "Istat", y = "medv", title = "Scatterplot medv v/s Istat
with nox")+scale_shape_identity()
g3
```



Here we concluded that nitrogen oxides concentration in lower class area is less. we can say that lower class people live in less pollution area.